Pacific Symposium on Biocomputing 6:89-100 (2001)

THE PROTEIN NON-FOLDING PROBLEM: AMINO ACID DETERMINANTS OF INTRINSIC ORDER AND DISORDER

R.M. WILLIAMS, Z. OBRADOVIĆ
School of Electrical Engineering and Computer Science,
Washington State University, Pullman, WA 99164-2670

V. MATHURA, W. BRAUN
Department of Human Biological Chemistry and Genetics,
Sealy Center for Structural Biology
University of Texas Medical Branch, Galveston 77555-1157

E.C. GARNER, J. YOUNG, S. TAKAYAMA,
C.J. BROWN, A.K. DUNKER
dunker@disorder.chem.wsu.edu
School of Molecular Biosciences,
Washington State University, Pullman, WA 99164-4660

To investigate the determinants of protein order and disorder, three primary and one derivative database of intrinsically disordered proteins were compiled. The segments in each primary database were characterized by one of the following: X-ray crystallography, nuclear magnetic resonance (NMR), or circular dichroism (CD). The derivative database was based on homology. The three primary disordered databases have a combined total of 157 proteins or segments of length ≥ 30 with 18,010 residues, while the derivative database contains 572 proteins from 32 families with 52,688 putatively disordered residues. For the four disordered databases, the amino acid compositions were compared with those from a database of ordered structure. Relative to the ordered protein, the intrinsically disordered segments in all four databases were significantly depleted in W, C, F, I, Y, V, L and N, significantly enriched in A, R, G, Q, S, P, E and K, and inconsistently different in H, M, T, and D, suggesting that the first set be called *order-promoting* and the second set *disorder-promoting*. Also, 265 amino acid properties were ranked by their ability to discriminate order and disorder and then pruned to remove the most highly correlated pairs. The 10 highest-ranking properties after pruning consisted of 2 residue contact scales, 4 hydrophobicity scales, 3 scales associated with β-sheets and one polarity scale. Using these 10 properties for comparisons of the 3 primary databases suggests that disorder in all 3 databases is very similar, but with those characterized by NMR and CD being the most similar, those by CD and X-ray being next, and those by NMR and X-ray being the least similar.

1 Introduction

1.1 The Extended Central Dogma of Molecular Biology

Information flow in molecular biology is generally taken to be:

DNA Sequence → RNA Sequence → AA Sequence → 3 D Structure → Function, where prediction of "*AA Sequence → 3 D Structure*" is called "the protein folding problem"[1] and where "*AA Sequence → 3 D Structure → Function*" is the generally accepted protein structure/function paradigm[2].

1.2 Intrinisic Disorder

In contradistinction to the scheme given above, many protein segments[3,4] and a few whole proteins[5-7] don't fold under their putative physiological conditions and yet exhibit function. The existence of such intrinsic disorder has led to a call for the re-assessment of the protein structure function paradigm.[8]

Recognizing the over-simplification of the partition into two states, order and disorder, and recognizing that all protein structure is condition-dependent, we are nevertheless focusing on an admittedly simplified problem: the prediction of intrinsic order and disorder from amino acid sequence,[9-14] or what we herein call "*the protein non-folding problem*." Application of our predictors to sequence databases suggests that there is a high amount of intrinsic disorder, with perhaps more than 25% of all proteins having disordered regions of 40 residues or longer.[15]

Here we report a substantial enlargement of our databases of intrinsic protein disorder and comparisons of these with ordered protein structure. The results provide insight into amino acid sequence features that determine whether a segment is likely to be intrinsically ordered or disordered. These new insights should lead to improved predictions of disorder and to an improved classification of types, or "flavors"[11,14] of disorder.

2 Materials and Methods

2.1 Databases

As before,[9,12] residues with missing backbone coordinates the Protein Data Bank (PDB)[16] were classified as disordered. We used PDB_Select_25,[(ref17)] rather than PDB itself to avoid redundancy. Segments and proteins characterized as disordered by nuclear magnetic resonance (NMR), by circular dichroism (CD) or protease sensitivity (PS) were located by key word searches using PUBMED, and the specifications of order and disorder made by the authors were used.

A derivative database was developed using sequence homology. BLAST searches[18] were followed by ClustalW[19] to find and align sequences related to an arbitrary group of 32 disordered proteins or segments, with 12 characterized by X-ray diffraction, 14 by NMR, 5 by CD and 1 by PS. Putative regions of disorder were identified by their homology to the known region of disorder.

2.2 Comparison of Amino Acid Compositions

To compare amino acid compositions of a specific disordered database, *a*, with those of the ordered database, *b*, the following was computed for each amino acid,

$$(M_j^a - M_j^b) / M_j^b \quad (1)$$

where M_j^a is the mole fraction of amino acid j in the disordered database, *a*, and M_j^b is the mole fraction of this same amino acid in the ordered data base.

From statistics,[20] the variances for these ratios are:

$$Var\ (M_j^a - M_j^b)/M_j^b = (M_j^a/M_j^b)^2 \{Var(M_j^a)/(Mj^a)^2 + Var(M_j^b)/(M_j^b)^2\}, \quad (2)$$

where M_j^a and M_j^b are as before and where $Var(M_j^a)$ and $Var(M_j^b)$ are the variances of amino acid *j* for databases *a* and *b*, respectively, and where, the standard deviation is the square root of the variance..

2.3 Amino Acid Properties

The values for various amino properties such as hydropathy, polarity, volume, etc., were compiled by database and literature searches. Altogether 265 distinct property scales were found, but many of these scales are highly correlated with each other. These 265 scales along with a matrix of correlations coefficients are available on our website: disorder.chem..wsu.edu.

2.4 Comparison of Amino Acid Properties

Using balanced numbers of ordered and disordered segments, plots of the conditional probabilities of order and disorder versus the property values were constructed for each of the 265 properties. The properties were then ranked by the relative degree of separation of the two probability curves using the area ratio method described in more detail previously.[21,22] Next, the correlation coefficients were calculated for the highly ranked properties; for sets with pair-wise correlation coefficients ≥ 0.9, only the highest-ranking property was kept.

3 Results

3.1 Databases of ordered and disordered proteins and segments

The database of ordered structure, called *O_PDB_Select 25*, and the four databases of disordered structure, called *dis_X-ray, dis_NMR, dis_CD and dis_Fam32*, are summarized in Table 1. An additional database, called *dis_ALL*, is the union of the three primary databases. The sequences and identities of the ordered and disordered residues for the proteins and segments in these databases can be found on our website: disorder.chem.wsu.edu.

Table 1. Data Summary

Group	Number of Segments	Number of Residues
O_PDB_Select_25	1,111	220,668
dis_X-ray	59	3,907
dis_NMR	43	4,108
dis_CD	55	10,818
dis_ALL	157	18,001
dis_Fam32	572	52,688

3.2 Comparison of the Amino Acid Compositions

In order to compare each of the four databases of disordered protein, *a*, with the ordered database, *b*, the fractional difference, $(M_j^a - M_j^b) / M_j^b$, was determined for each amino j (Figure 1). Thus, a negative peak for amino acid j indicates that the given disordered database is depleted in that amino acid compared to the ordered dataset, while a positive peak indicates enrichment. The amino acids in this figure were arranged by Vihinen et al.'s flexibility index,[23] which is inferred to relate to the tendency of a given amino acid type to be buried (left) or exposed (right).

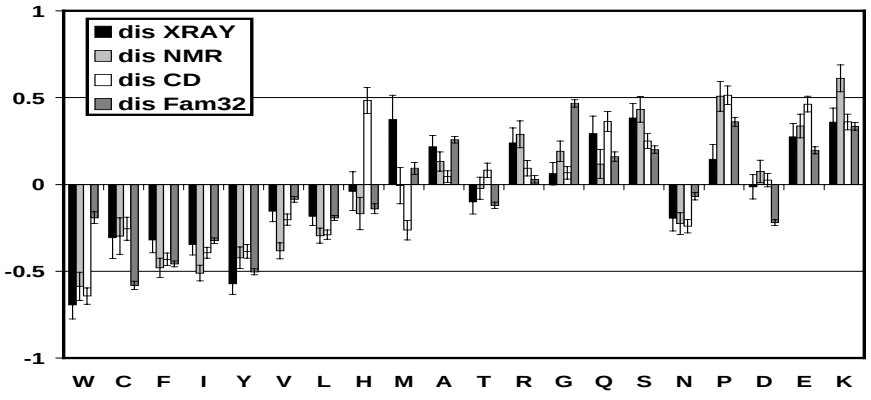

Figure 1. Amino acid composition of each database relative to the ordered dataset.

For all four databases, disordered regions are consistently depleted in W, C, F, I, Y, V, L, and N, and consistently enriched in A, R, G, Q, S, P, E, and K. As indicated by the error bars, most of these enrichments and depletions are greater than 3 standard deviations from the value for ordered proteins.

3.3 Comparison of Amino Acid Properties

The disordered residues in dis_ALL, were balanced by an equal number of ordered residues selected at random from the ordered database. From this balanced data and with an averaging of the property values over windows of 21 residues, the conditional probabilities of order and disorder were determined and plotted versus the property values as described previously.[22] This procedure was repeated 5 times, with 5 random selections of ordered protein without replacement. The resulting 5 ordered and 5 disordered curves for the 14 Å contact number of Nishikawa and Ooi[24] are shown in Figure 2. This property measures the exposure of a residue to the solvent, and is related closely to the distance from the center of mass of a protein. It is defined as the number of C alpha atoms surrounding the residue located within a sphere of the radius of 14 Å, and is derived from a statistical analysis of residues in proteins with known 3D structure.

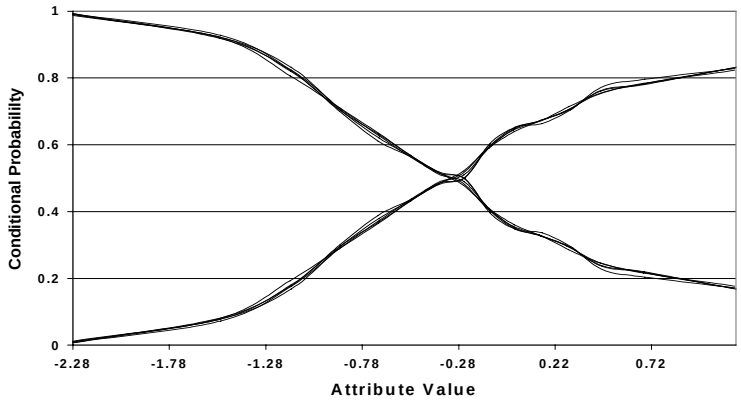

Figure 2. Conditional probability curves for 5 ordered datasets vs. dis_ALL for the 14 Å Contact Number

The greater the separation of the two curves, the better a given property distinguishes between the order and disorder of the input data.[22] This separation can be quantified by dividing the area bounded by the two curves by the total area to give the area ratio (AR). Application of this procedure to the 5 pairs of curves in Figure 2 gives AR values of 0.536, 0.533, 0.535, 0.543, and 0.538 or an <AR> of

0.537 ± 0.004. Thus, the AR is insensitive to the randomly selected sets of order used in the analysis.

The conditional probability curves are shown in Figure 3 for this same property but with the three primary databases of disordered proteins used individually. The different databases of disorder show similar but distinguishable curves for this attribute. Each of these curves was constructed 5 times with different collections of ordered segments as before, with the resulting AR values of 0.424 ± 0.014 for dis_X-ray, 0.605 ± 0.004 for dis_NMR, and 0.540 ± 0.003 for dis_CD.

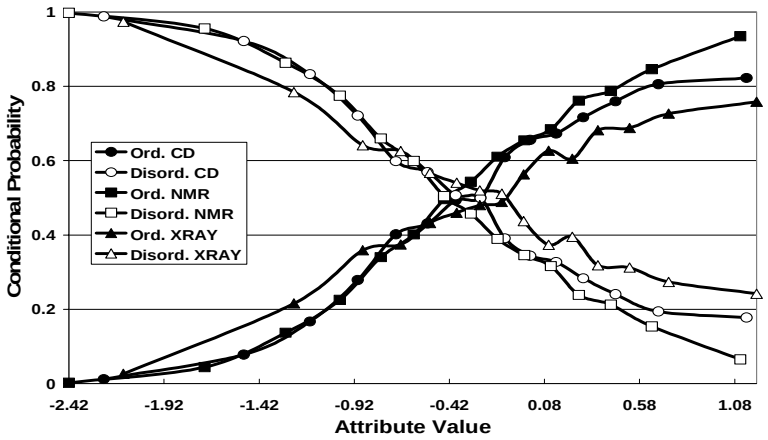

Figure 3. Conditional probability curves for each of the disordered databases for the 14 Å Contact Number

Using dis_ALL as for Figure 1, this same procedure was repeated for each of the 265 property scales and the resulting AR values were used to rank-order the properties. Upon removing properties with correlation coefficients ≥ 0.9, the top 40 reduced to a set of 10. These 10 properties are given in Table 2 along with their AR values and rankings for each of the four disorder databases. For the most part, rankings were similar for the different databases, for example between 1 and 9 for the first property, 3 and 13, for the second property, etc.

Correlation coefficients (values of r) and levels of significance (p-values) for the 10 properties are given as a matrix (Table 3), with the r-values above the diagonal and the p-values below. These 10 properties exhibit a range of correlation coefficients, from 0.604 to 0.894, with an overall average for their absolute values of 0.76 ± 0.09.

Table 2. Selected properties that distinguish intrinsic order and disorder.

	PROPERTY	Ref	ALL	CD	NMR	X-ray	Fam32
1	14 Å Contact Number		0.537 [1]	0.540 [2]	0.605 [2]	0.424 [9]	0.424 [8]
		24	0.004	0.003	0.004	0.014	0.003
2	Optimal matching hydrophobicity		0.528 [3]	0.538 [4]	0.588 [9]	0.418 [13]	0.424 [9]
		26	0.003	0.004	0.004	0.015	0.001
3	Beta sheet propensity		0.510 [19]	0.504 [15]	0.612 [1]	0.406 [20]	0.396 [20]
		27	0.003	0.006	0.005	0.011	0.003
4	HPLC Hydrophobicity		0.494 [21]	0.501 [17]	0.534 [41]	0.438 [4]	0.370 [36]
		28	0.002	0.003	0.004	0.013	0.003
5	Hydrophobic parameter pi		0.489 [24]	0.493 [21]	0.529 [42]	0.415 [15]	0.355 [56]
		29	0.002	0.005	0.005	0.012	0.001
6	Fraction of site occupied by water		0.476 [35]	0.478 [32]	0.526 [46]	0.366 [45]	0.394 [21]
		30	0.003	0.005	0.007	0.015	0.003
7	Information measure for pleated sheet		0.478 [33]	0.482 [29]	0.560 [27]	0.326 [79]	0.373 [30]
		31	0.003	0.007	0.005	0.011	0.002
8	Partition free energy		0.481 [30]	0.471 [42]	0.563 [24]	0.410 [18]	0.381 [25]
		32	0.003	0.004	0.005	0.016	0.002
9	Coordination number		0.476 [34]	0.464 [50]	0.563 [25]	0.418 [14]	0.438 [5]
		33	0.003	0.001	0.006	0.007	0.002
10	Free-energy beta-strand conformation		0.472 [40]	0.467 [49]	0.568 [21]	0.356 [53]	0.359 [45]
		34	0.002	0.006	0.005	0.014	0.002

Table 3. Correlation coefficients (above) and p-values (below) among properties.

	1	2	3	4	5	6	7	8	9	10
1	1	0.865	0.809	0.841	0.893	-0.855	0.810	-0.877	0.816	-0.840
2	0.0001	1	0.809	0.825	0.847	-0.687	0.799	-0.798	0.628	-0.802
3	0.0001	0.0001	1	0.691	0.720	-0.633	0.894	-0.701	0.705	-0.888
4	0.0001	0.0001	0.0007	1	0.873	-0.679	0.604	-0.746	0.662	-0.633
5	0.0001	0.0001	0.0003	0.0001	1	-0.758	0.660	-0.847	0.609	-0.674
6	0.0001	0.0008	0.0028	0.001	0.0001	1	-0.659	0.864	-0.694	0.695
7	0.0001	0.0001	0.0001	0.0048	0.0015	0.0016	1	-0.674	0.685	-0.888
8	0.0001	0.0001	0.0006	0.0002	0.0001	0.0001	0.0011	1	-0.708	0.714
9	0.0001	0.003	0.0005	0.0015	0.0043	0.0007	0.0009	0.0005	1	-0.755
10	0.0001	0.0001	0.0001	0.0028	0.0011	0.0007	0.0001	0.0004	0.0001	1

4 Discussion

4.1 The Input Data

Compared to our previously published studies,[9,11] the number of disordered residues reported here is increased by more than 100-fold. The largest amount of primary data is from characterization by CD, which is only semi-quantitative and which lacks position-specific information. However, these disordered data are similar to those obtained by the other methods of characterization.

Use of homology provides a means to rapidly increase the amount of disorder data (Table 1) An argument against use of homology is that such sequences are correlated and so adding sequences by this method does not increase the information content very effectively. However, for many proteins with disordered regions, the disordered parts show significantly less sequence similarity than do the ordered parts (work in progress), suggesting that identification of disordered regions by homology is apparently an effective way to increase the information content after all. An additional problem is that a corresponding region could be disordered in one protein but ordered in its homologue as observed for proteins in the prion family.[35]

4.2 Amino Acid Compositions

The flexibility index used to specify the arrangement of the amino acids in Figure 1 was based on the B-factor values of the backbone atoms associated with each residue type, averaged over 92 unrelated proteins.[23] These values are determined less by intrinsic flexibility and more by the tendency of a given amino acid type to be buried (left) or exposed (right). Thus, the ordered proteins contain a higher proportion of amino acids that tend to be buried, while disordered proteins have a higher proportion of amino acids that tend to be on the surface of ordered proteins.

For the ordered database, 45% of the amino acids are from the W to A set (e.g. the 10 left-most), with 55% from T – K (e.g., the 10 right-most). For dis_ALL, the left-most 10 comprise 34% and the 10 right-most 66%, while for dis_Fam32 the corresponding numbers are 37% and 63%. Thus, the balance of order-promoting and disorder-promoting amino acids correlates with whether a protein or segment is intrinsically ordered or disordered.

Disordered segments are not substantially enriched in T, N and D as expected from the behaviors of the neighboring amino acids in Figure 1. We speculate that the anomalous behavior of these three amino acids results from their polar β-carbon branches, which can form hydrogen bonds with the peptide group. These hydrogen bonds would lower the configurational entropy of the backbone in the disordered state and thereby reduce the promotion of disorder by T, N and D.

4.3 Amino Acid Properties

We sought to identify a set of 10 attributes that were ranked in the top 15% for discriminating order and disorder and at the same time were correlated as little as possible with each other. Meeting the first criterion meant that only the top 40 were considered, e.g. 40/265 = 15%. Meeting the second criterion led to a pruning cut-off of 0.9 for the correlation coefficient, which reduced the group from 40 to 10.

The flexibility index was used for Figure 1 because of our prior experience that this property gave good discrimination between segments of order and disorder [9,25] and because a scale based on flexibility is easy to explain for this purpose. However this scale does not appear in Table 2. Its absence is not due to a poor discrimination between order and disorder, however: this scale ranks 9, 13, 3, and 3 for dis_ALL, dis_CD, dis_NMR, and dis_X-ray, respectively. This scale does not appear in Table 2 because it has a 0.95 correlation coefficient with the top-ranked property and so was lost in the pruning process.

The top-ranked property for distinguishing order and disorder for the dis_ALL database is the 14 Å contact number[24] and so was used for illustration in Figures 2 and 3. Note that a related property, the coordination number, also ranks high, at position 9 in Table 2 and at number 34 overall. Note also that these two scales have a correlation coefficient of 0.865 with a level of significance of 0.001. Both of these properties relate to the number of side chains found close to a given side chain in a set of proteins of known structure. In a sense, these numbers provide a ranking of the ability of the various amino acids to be tightly packed. It is interesting that a measure of "packing capacity", not hydropathy or net charge, ranks first for discriminating order and disorder for the data currently on hand.

Of the 10 properties in Table 2, four are associated with hydrophobicity: entries 2, 4, 5, and 8; and one is associated with polarity: entry 6. Of special note is that the Kyte and Doolittle scale[36], which is perhaps the most widely used scale of this type, ranked below the 5 similar scales in Table 2. That is, the Kyte and Doolittle scale gave an AR value of 0.420 and ranked 79th for dis_ALL, while the 5 similar scales in Table 2 gave values ranging from 0.528 to 0.476 and rankings from 3 to 35. From the perspective of protein folding, the hydrophobicity scale of Sweet and Eisenberg[26] is evidently the best so far.

The remaining 3 properties at positions 3, 7 and 10 in Table 2 all relate to the propensity of amino acids to form β–strands. There may be a simple structural explanation for this result. Amino acids with branches at the β-carbon both reduce flexibility and favor β-strand formation, suggesting that the two scales should be negatively correlated. Indeed, the β-strand propensities (number 3 in Table 2) and the flexibility index of Figure 1 have a correlation coefficient of – 0.83. In addition, there might be biological selection against intrinsically disordered regions with a high propensity to form sheets. Such sequences could be expected to have a

tendency to form amyloid-type β-polymers. To test this possibility, we studied the prion sequence. The region identified as crucial to polymer formation has a segment of disorder that also has a high propensity for β-sheet. Further studies of this issue are clearly warranted.

4.4 Similarity of Disorder Characterized by Different Methods

The regions of intrinsic disorder characterized by different methods exhibit similar amino acid compositions (Figure 1). Differences in AR values (Table 2) provide a second way to compare pairs of databases. For this comparison, the absolute values of the AR differences were calculated for the 10 properties in Table 1 and averaged for pairs of databases: CD / NMR, 0.07 ± 0.03; CD / X-ray, 0.10 ± 0.03 and NMR / X-ray, 0.17 ± 0.04. By this measure, disordered sequences characterized by CD and NMR are most similar to each other, those by CD and X-ray are next in similarity, and those by NMR and X-ray are the least similar, although the standard deviations indicate that the overall differences are small. NMR and CD might yield the most similar disorder because almost all of the proteins in both sets are likely to be fully unfolded (e.g. random coil-like), while X-ray might yield a slightly different type of disorder because a significant proportion of the protein in this set could be partially folded (e.g. molten globule-like).

4.5 Work in Progress and Future Directions

The increased sizes of our disordered databases and the increased understanding of the sequence determinants of order and disorder are enabling us to identify different types or flavors of disorder. When predictions are carried out on a flavor-by-flavor basis, accuracies seem to improve. The next goal will be to determine whether there are relationships between disorder flavor and protein function.

Acknowledgments

Support from NSF-CSE-IIS-9711532 to Z. O. and A. K. D., from NSF REU IID 9711532 to R. M. W. and Z. O., from N.I.H. 1R01 LM06916 to A.K.D and Z.O. and from DOE DE-FG03-96ER62267 and from the Texas Higher Education Coordinating Board ARP 004952-0084-1999 to W.B. are gratefully acknowledged.

References

1. Creighton TE,. “The protein folding problem” *Science* **240**, 267- 344 (1988)

2. Orengo CA, Todd AE, “From protein structure to function”. *Curr. Opin. Struct.* Biol. **3**, 374-382 (1999)
3. Spolar RS, Record II MT. “Coupling of local folding to site-specific binding of proteins to DNA” *Science* **263**, 777-784 (1994)
4. Plaxco KW, Gross M, “The importance of being unfolded” *Nature* **386**, 657, 659 (1997)
5. Daughdrill GW, Chadsey MS, Karlinsey JE, Hughes KT, Dahlquist FW, “The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28”. *Nat. Struct. Biol.* **4**, 285-291 (1997)
6. Weinreb PH, Zhen W, Poon AW, Conway KA, Lansbury PT, Jr, “NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded” *Biochemistry* **35**, 13709-13715 (1997)
7. Fletcher CM, McGuire AM, Gingras AC, Li H, Matsuo H, Sonenberg N, Wagner G, “4E binding proteins inhibit the translation factor eIF4E without folded structure” *Biochemistry* **37**, 9-15 (1998)
8. Wright PE, Dyson HJ, “Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm” *J. Mol. Biol.* **293**, 321-331 (1999)
9. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Dunker AK, “Identifying disordered regions in proteins from amino acid sequences” *Proc. I.E.E.E. Internat. Conf. Neural Networks* **1**, 90-95 (1997)
10. Romero P, Obradovic Z, Dunker AK, “Sequence data analysis for long disordered regions prediction in the calcineurin family” *Genome Informatics* **8**, 110-124 (1997)
11. Romero PZ, Obradovic C, Dunker AK,. “Intelligent data analysis for protein disorder prediction” *Artificial Intelligence Reviews*: In Press (2000)
12. Li X, Romero P, Rani M, Dunker AK, Obradovic Z, ”Predicting protein disorder for N-, C-, and internal regions” *Genome Informatics* **10**, 30-40 (1999)
13. Garner E, Cannon P, Romero P, Obradovic Z, Dunker AK, “Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization” *Genome Informatics* **9**, 201-213 (1998)
14. Garner E, Romero P, Dunker AK, Brown C, Obradovic Z, “Predicting binding regions within disordered proteins” *Genome Informatics* **10**, 41-50 (1999)
15. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Guilliot S, Garner E, Dunker AK, “Thousands of proteins likely to have long disordered regions” *Pacific Symp. Biocomput.* **3**, 437-448 (1998)
16. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, “The protein data bank” *Nucleic Acids Res.* **28**, 235-242 (2000)
17. Hobohm U, Scharf M, Schneider R, Sander C, “Selection of representative protein data sets” *Protein Sci.* **1**, 409-417 (1992)
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, “Basic local alignment search tool” *J. Mol. Biol.* **215**, 403-410 (1990)

19. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ, "Multiple sequence alignment with Clustal X" *Trends Biochem Sci* **23**, 403-405 (1998)
20. Kendall M, Stuart A. *The Advanced Theory of Statistics*: Charles Griffin & Company Limited; 1977. 455 p.
21. Arnold GE, Dunker AK, Johns SJ, Douthart RJ, "Use of conditional probabilities for determining relationships between amino acid sequence and protein secondary structure" *Proteins: Structure, Function and Genetics* **12**, 382-399 (1992)
22. Xie Q, Arnold GE, Romero P, Obradovic Z, Garner E, Dunker AK, "The sequence attribute method for determining relationships between sequence and protein disorder" *Genome Informatics* **9**, 193-200 (1998)
23. Vihinen M, Torkkila E, Riikonen P, "Accuracy of protein flexibility predictions" *Proteins* **19**, 141-149 (1994)
24. Nishikawa K, Ooi T, "Radial locations of amino acid residues in a globular protein: correlation with the sequence" *J Biochem* (Tokyo) **100**, 1043-1047 (1986)
25. Dunker A, Obradovic Z, Romero P, Kissinger C, Villafranca E, "On the importance of being disordered" *PDB Newsletter* **81**, 3-5 (1997)
26. Sweet RM, Eisenberg D, "Correlation of sequence hydrophobicities measures similarity in three- dimensional protein structure" *J. Mol. Biol*. **171**, 479-488 (1983)
27. Palau J, Argos P, Puigdomenech P, "Protein secondary structure. Studies on the limits of prediction accuracy" *Int. J. Pept. Protein Res*. **19**, 394-401 (1982)
28. Wilson KJ, Honegger A, Stotzel RP, Hughes GJ, "The behaviour of peptides on reverse-phase supports during high-pressure liquid chromatography" *Biochem. J*. **199**, 31-41 (1981)
29. Fauchere JL, Pliska V,. *Eur. J. Med. Chem*. **18**, 369-375 (1983)
30. Krigbaum WR, Komoriya A, "Local interactions as a structure determinant for protein molecules: II" *Biochim Biophys Acta* **576**, 204-248 (1979)
31. Robson B, Suzuki E, "Conformational properties of amino acid residues in globular proteins" *J. Mol. Biol.* **107**, 327-356 (1976)
32. Guy HR, "Amino acid side-chain partition energies and distribution of residues in soluble proteins" *Biophys. J*. **47**, 61-70 (1985)
33. Galaktionov SG, Marshall GR, *Prediction of protein structure in terms of intraglobular contacts: 1D to 2D to 3D* St. Louis, MO. Washington University Institute for Biomedical Computing (1996)
34. Munoz V, Serrano L, "Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales" *Proteins* **20**, 301-311 (1994)
35. Marcotte, EM, Eisenberg, D, "Chicken prion tandem repeats form a stable, protease-resistant domain" *Biochemistry* **38**, 667-676 (1999)
36. Kyte J, Doolittle RF, "A simple method for displaying the hydropathic character of a protein" *J. Mol. Biol*. **157**, 105-132 (1982)